

# Supplementary Materials to Establishing the reliability and validity of measures extracted from long-form recordings

## Contents

About this document . . . . .	1
Read in all data . . . . .	2
SM A: Simulation to better understand ICC . . . . .	2
SM B: More information for benchmarking our results against previously reported reliability studies	3
SM C: Code to reproduce Table 2 . . . . .	4
SM D: Code to reproduce Fig. 1 . . . . .	4
SM E: Code to reproduce text at the beginning of the “Setting the stage” section . . . . .	6
SM F: Code to reproduce Table 3 . . . . .	6
SM G: Code to reproduce text above Figure 4 . . . . .	6
SM H: Code to reproduce Figure 4 . . . . .	6
SM I: Is lower Child ICC than correlations due to the fact that we are controlling for age? (Exploration)	6
SM J: Code to reproduce text at the beginning of the “Overall reliability” section . . . . .	7
SM K: Are high Child ICCs for “other child” measures due to number or presence of siblings? (Exploration) . . . . .	7
SM L: Are “bad” output measures those coming from VCM? (Exploration) . . . . .	8
SM M: Code to reproduce Figure 5 . . . . .	8
SM N: Code to reproduce text below Figure 5 . . . . .	8
SM O: Code to reproduce Table 4 . . . . .	10
SM P: Code to reproduce text at the beginning of the “Reliability across age groups” section . . .	10
SM Q: Code to reproduce Figure 6 . . . . .	10
SM R: Code to reproduce text below Figure 6 . . . . .	10
SM S: Code to reproduce Figure 7 . . . . .	10
SM T: Code to reproduce text in the “Reliability within corpus” section . . . . .	10
SM U: Code to reproduce Figure 8 . . . . .	13
SM V: Code to reproduce text below Figure 8 . . . . .	13
SM W: Code to reproduce Figure 9 . . . . .	13
SM X: Code to reproduce text in the Discussion section . . . . .	13
SM Y: Variability as a function of hardware . . . . .	15
SM Z: References . . . . .	15

## About this document

The pdf version of this document was generated using an RMarkdown document, which includes all the computer code to reproduce key figures, tables, and analyses mentioned in the main manuscript, as well as additional texts and analyses. In the pdf version, you will see two types of text:

Quoted text indicates sections that appear verbatim in the main manuscript.

In contrast, unmarked text are additional texts and analyses, as well as comments for the reader who is attempting to reproduce our analyses. Readers attempting to reproduce our analyses are asked to also read comments in chunks of code in the RMarkdown version of the document, which are invisible in the pdf version. This is indicated with “Please see code in the RMarkdown version of the document.”

## Read in all data

### SM A: Simulation to better understand ICC

We were uncertain of how to interpret ICC's numeric values. It is described as “proportion of variance explained”, but we do not know if it should be considered as a percentage (like  $R^2$ ) or a correlation (like  $r$ ). We therefore simulated data controlling the underlying  $r$  between paired datapoints to see how ICC recovered that underlying  $r$ .

We informed the simulation by the data we have as follows:

- we had the same number of corpora, and of children in each corpus
- we had the same metrics for each (i.e., CVC, AWC, etc) – and these metrics had the same mean & SD for day 1 of recordings as in observed data

In the simulation, we departed from reality as follows:

- we did not consider the  $r$  across multiple days observed in the data, but instead generate data points to vary  $r$  from a small correlation ( $r=.1$ ), a moderate one ( $r=.3$ ), a larger one ( $r=.5$ ). It is unlikely that test-retest correlations in infancy would be much greater than .5 (based on previous studies using test-retest in experimental tasks), but for completeness, we also use .7 and .9
- we did not consider child age, nor variable re-recording periods
- we had a single pair of recordings (rather than variable number of re-recordings)

We used `simstudy`, a package created for such simulations, following the vignette <https://cran.r-project.org/web/packages/simstudy/vignettes/correlated.html> to create correlated data providing a correlation matrix. Results are shown in the Figure below. Each point represents the ICC extracted from a mixed model applied to one metric, combining data from all corpora. It appears that ICC values reflect underlying  $r$  values, but underestimating  $r$  more the larger  $r$  is.

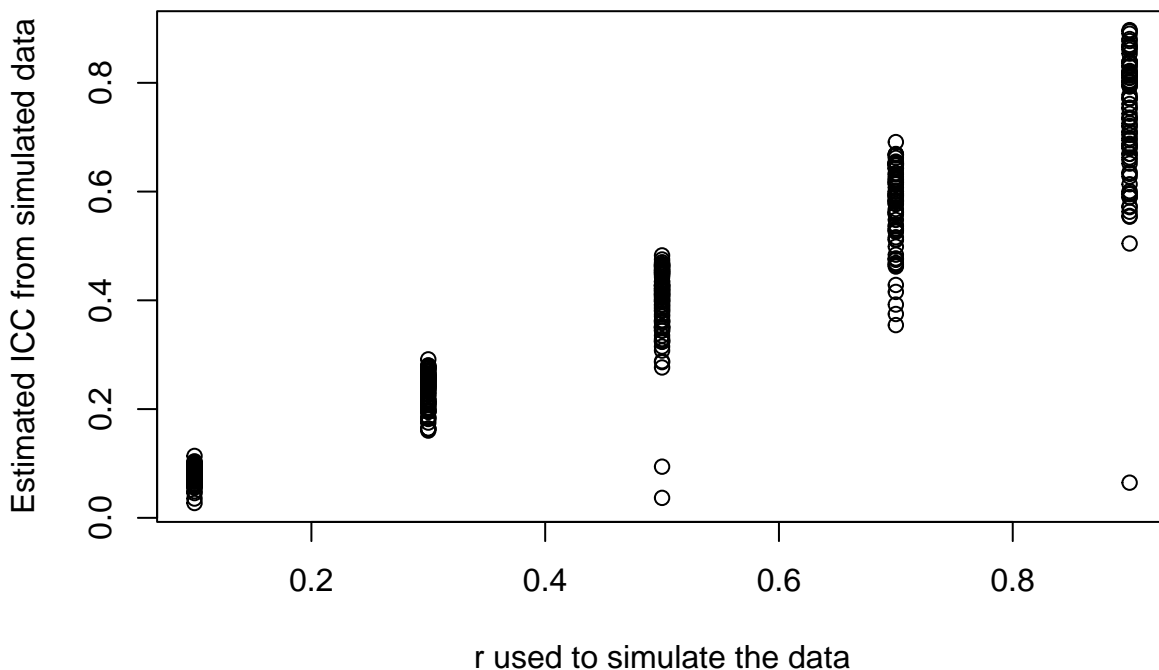


Figure 1: Results from simulations aimed at helping understand the relationship between underlying  $r$  and estimated ICC.

## SM B: More information for benchmarking our results against previously reported reliability studies

First, we looked for measures of language development used with observational data that can be employed with children aged 0-3 years, and which are available at least in English. To this end, we did searches in scholar.google.com and google.com with keywords like (“language development” AND “standardized test” AND “infancy”) around February 2021, complementing these searches with our own knowledge.

All of the instruments we found rely on reports from caregivers, including the Child Observation Record Advantage (Schweinhart, McNair, & Lerner, 1993); the Desired Results Developmental Profile (Kriener-Althen, Newton, Draney, & Mangione, 2020); and the MacArthur-Bates Communicative Development Inventory (MB-CDI, Fenson et al., 1994). Readers are likely most familiar with the MB-CDI, so we focus our discussion on this instrument here, to illustrate the level of test-retest reliability such instruments can attain.

In the work most commonly cited in relation to the MB-CDI’s creation, Fenson et al. (1994) reported a correlation of  $r=.95$  in their sample of North American, monolingual infants. Since it has been over 20 years, and this instrument has been used with other samples and languages, a more relevant comparison point for our own study would be one that takes into account more attempts at using this instrument. We did not find a systematic review or meta-analysis on test-retest estimates for the MB-CDI. However, Frank et al. (2021) analyzed data archived in an MB-CDI repository, concentrating on two corpora where longitudinal data was available, American English and Norwegian. They found that, for both datasets, correlations within 2-4 months were above  $r=.8$ , with correlations at 16 months of distance (i.e., a form filled in at 8 and 24 months) at their lowest,  $r=.5$ . These correlations are very high when considering that the MB-CDI tends to have ceiling and floor effects at these extreme ages. Alcock, Meints, and Rowland (n.d.) looked at correlations when British English-speaking parents completed two versions of the form in two media (e.g., short form in paper and long form online, or vice versa) within a month. Here, the correlation was  $r=.8$  for comprehension and  $r=.6$  for production.

Thus, although evidence is scant, it is sufficient to see that the MB-CDI can lead to high correlations, that are nearly comparable to those expected in standardized testing among adults. It is worth bearing in mind that test-retest reliability in parental report measures does not depend only on the consistency in the individual infants’ relative scores for a given behavior, but also on the consistency of the adult in reporting it. Moreover, they are based on cumulative experience, rather than a one-shot observation, as in the case of long-form recordings. Therefore, they do not constitute an appropriate comparison point, and by and large we can be quite certain that they will yield higher reliability than metrics based on the children themselves. Thus, parental report may not be the most appropriate comparisons for our own study because they may overestimate reliability in the child’s own behavior.

Second, we did a bibliographic search for systematic reviews of test-retest reliability of standardized instruments to measure language development up to age three years that are available at least in English using similar methods to those mentioned above, with the searches performed in June 2023. Although reliability in these tests may also partially reflect consistency in the adults’ reports, they are at least based on a one-shot observation of how the child behaves, rather than the adult’s cumulative experience with the child. Note that some promising tests are currently in development or have only recently been released, including the NIH Baby Toolbox (NIH, 2023) and GSED (Cavallera et al., 2023), and thus we cannot include them in the present summary.

The Ages and Stages Questionnaire (ASQ) is a screening tool to measure infants and children’s development based on observation of age-specific behaviors: For instance, at 6 months, in the domain of communication, one of the items asks whether the child smiles. The ASQ’s reliability has been the object of a systematic review of independent evidence (Velikonja et al., 2017). Across 10 articles with data from children in the USA, Canada, China, Brazil, India, Chile, the Netherlands, Korea, and Turkey, only three articles reported test-retest correlations, two in the USA ( $r=.84-1$ ) and one in Turkey ( $r=.67$ ). However, the meta-analysis authors judged these three studies to be “poor” in quality. Moreover, the ASQ is a questionnaire that an observer fills in, but it can also be administered as a parental questionnaire, reducing its comparability with our purely observational method.

For the other tests, reliability is mainly available from reports by the companies commercializing the test. The Goldman Fristoe Articulation Test – 3rd edition (GFTA-3) (Goldman & Fristoe, 2015) focuses on the ability to articulate certain sounds through picture-based elicitation and can be used from two to five years of age. Available in English and Spanish for a USA context, it has a reported reliability of  $r=.92$  – although we do not know whether it is this high for two-year-olds specifically. The Preschool Language Scales – 5th edition (Zimmerman, Steiner, & Pond, 2011) can be used to measure both comprehension and production from birth to 7 years of age. According to one report (Leaders Project, 2013), its test-retest reliability is .86 to .95, depending on the age bracket (0;0-2;11, 3;0-4;11, and 5;0-7;11, 0-7 years). The test has also been adapted to other languages, with good reported test-retest reliability ( $r=.93$ ; Turkish, Sahli & Belgin, 2017). One issue we see with both of these reports is that children tested varied greatly in age, and the correlation seems to have been calculated based on the raw (rather than the normed) score. As a result, children’s relative scores may have been stable over test and retest mainly because they varied greatly in age, rather than capturing variance from more meaningful individual differences. The Expressive Vocabulary Test – 2nd Edition (Williams, 1997) is a picture-based elicitation method that can be used with children from 2.5 years of age. The company Springer reports a test-retest reliability of  $r=.95$  by age (Gurley, 2011).

Many other standardized tests exist for children over three years of age. Given that most children included in the present study were under three, these other tests do not constitute an ideal comparison point. Nonetheless, in the spirit of informativeness, it is useful to consider that a systematic review and meta-analysis has been done looking at all psychometric properties (internal consistency, reliability, measurement error, content and structural validity, convergent and discriminant validity) for standardized assessments targeted at children aged 4-12 years (Denman et al., 2017). Out of 76 assessments found in the literature, only 15 could be evaluated for their psychometric properties, and 14 reported on reliability based on independent evidence (i.e., the researchers tested and retested children, rather than relying on the company’s report of reliability). Among these, correlations for test-retest reliability averaged  $r=.67$ , with a range from .35 to .76. The authors concluded that psychometric quality was limited for all assessments, but based on the available evidence, PLS-5 (whose test-retest reliability was  $r=.69$ ) was among those recommended for use.

Third, and perhaps most relevant, we looked for references that evaluated the psychometric properties of measures extracted from wearable data, also in June 2023. We found no previous work attempting to do so on the basis of completely ecological, unconstrained data like ours. The closest references we could find reported on reliability and/or validity of measurements from wearable data collected in constrained situations, such as having 4.5 year old children wear interior sensors and asking them to complete four tests of balance (e.g., standing with their eyes closed; Liu et al., 2022). It is likely that consistency and test-retest reliability are higher in such cases than in data like ours, making it hard to compare. Nonetheless, to give an idea, a recent meta-analysis of wearable inertial sensors in healthy adults (Kobsar et al., 2020) found correlations between these instruments and gold standards above  $r=.88$  for one set of measures (based on means) but much lower for another (based on variability, max weighted mean effect  $r=.58$ ). Regarding test-retest reliability, the meta-analysts report ICCs above .6 for all measures for which they could find multiple studies reporting them. However, those authors point out that the majority of the included studies were classified as low quality, according to a standardized quality assessment for that work.

## **SM C: Code to reproduce Table 2**

Please see code in the RMarkdown version of the document.

## **SM D: Code to reproduce Fig. 1**

Please see code in the RMarkdown version of the document.

A variable with one of the lowest Child ICC (LENA’s peak hourly AWC, Child ICC = 0.23) has a flatter regression line and wider confidence intervals, with points spread widely around the regression line. In contrast, the variable with the highest Child ICC (ACLEW’s other child vocalization duration, Child ICC = 0.63) has points that are less scattered and closer to the diagonal.

Table 1: Table 2 (reproduced).

corpus	location	chiXcor	recRXchi	recXcor	durXcor	ageXcor	ageRXcor
bergelson	Northeast US	44	10-12	522	14	11.2	6-17
cougar	Northwest US	26	3-45	239	11.1	26.6	0-59
fausey-trio	Western US	28	3-3	84	13.7	8.9	6-12
lucid	Northwest England	35	4-7	235	15.4	20	10-31
lyon	Central France	16	3-6	54	14.1	19.6	2-41
quechua	Highlands Bolivia	20	2-3	48	12.5	37.6	9-78
warlaumont	Western US	13	2-4	38	12.5	6.3	2-18
winnipeg	Western Canada	9	2-5	34	9.1	14.5	2-33

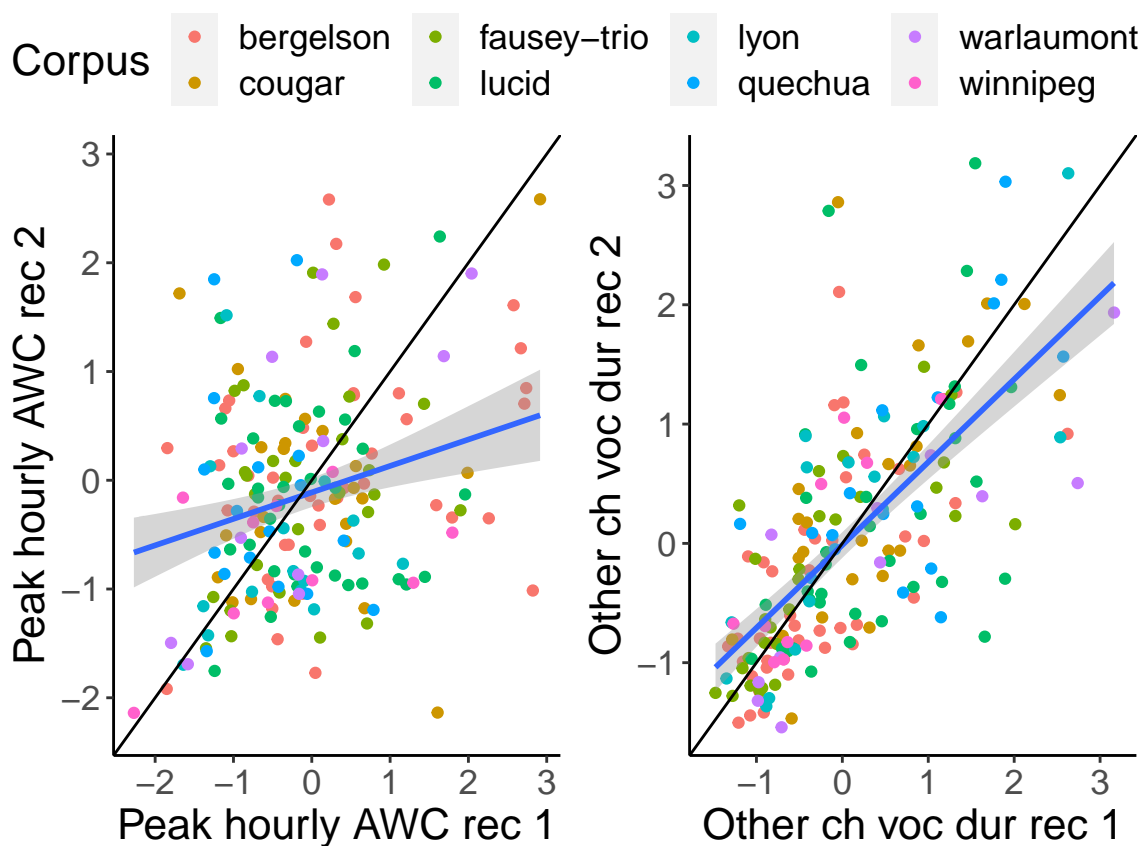


Figure 2: Figure 2 (reproduced). Scatterplots for two selected variables. The left one has relatively low ICCs; the right one has relatively higher ICCs.

Table 2: Table 3 (reproduced). Mean and range for Pearson correlations across sampled pairs of recordings

	aclew	lena
AWC	.55 [.47,.63]	.52 [.4,.64]
CVC	.8 [.76,.84]	.7 [.64,.76]
CTC	.76 [.72,.8]	.69 [.61,.77]
Chi vocs	.69 [.63,.75]	.61 [.53,.69]

Table 3: Type 3 ANOVA on model attempting to explain variation in Child ICC as a function of talker types and pipelines.

	Sum Sq	Df	F value	Pr(>F)
Type	0.36	4	7.42	0.00
data_set	0.04	1	3.28	0.07
Type:data_set	0.01	4	0.26	0.90
Residuals	0.78	65	NA	NA

### SM E: Code to reproduce text at the beginning of the “Setting the stage” section

Please see code in the RMarkdown version of the document.

Out of 191 children in 8 corpora, 148 children (belonging to 7 corpora) could be included in this analysis, as some children did not have recordings less than two months apart. In particular, no child from the Warlaumont corpus did.

### SM F: Code to reproduce Table 3

Please see code in the RMarkdown version of the document.

### SM G: Code to reproduce text above Figure 4

Please see code in the RMarkdown version of the document.

To see whether correlations in this analysis differed by talker types and pipelines, we fit a linear model with the formula  $lm(cor \sim type * pipeline)$ , where type indicates whether the measure pertained to the key child, (female/male) adults, other children; and pipeline LENA or ACLEW. The model was overall significant ( $F(65) = 3.74, p < .001$ ). We found an adjusted R-squared of 25%, suggesting this model did not explain a great deal of variance in correlation coefficients. A Type 3 ANOVA on this model revealed a significant effect of pipeline ( $F = 3.28, p = 0.07$ ), due to higher correlations for ACLEW ( $M = 0.57, SD = 0.13$ ) than for LENA metrics ( $m = M = 0.53, SD = 0.12$ ).

See table below for results of the Type 3 ANOVA.

### SM H: Code to reproduce Figure 4

Please see code in the RMarkdown version of the document.

### SM I: Is lower Child ICC than correlations due to the fact that we are controlling for age? (Exploration)

May it be that correlation coefficients are higher than Child ICC because we control for age in the latter but not the former? To assess this, we take a metric that shows a large difference across the correlation and the Child ICC analysis, ACLEW’s CVC per hour. We then refit our mixed model, but this time we do not

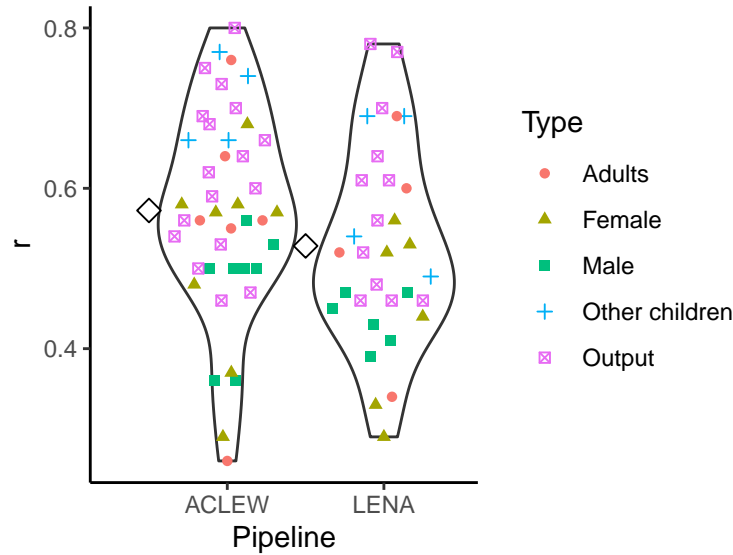


Figure 3: Figure 4 (reproduced). Violin plot reflecting the distribution of correlations.

declare age nor corpus, so that the situation is more similar to the simple correlations. When we do this, Child ICC goes up from 0.42 to 0.67.

This is still lower than the correlation observed for this same variable, 0.8. We believe this is due to another difference between the two analyses, namely that here we are including all recordings, whereas the other analysis is focused on recordings less than 2 months away.

### SM J: Code to reproduce text at the beginning of the “Overall reliability” section

Please see code in the RMarkdown version of the document.

Out of the 75 fitted models, 71 could be fit with the full model, yielding a measure of Corpus ICC. Of these, 97% had Corpus ICCs smaller than .2, consistent with the idea that LENA and ACLEW metrics are robust to corpus differences. For the 4 for which the full model was singular, we fit the data with the No Corpus model, and none was singular then, allowing us to have Child ICC for all 75 metrics.

Figure 5 shows the distribution of Child ICC across all 75 metrics, separately for each pipeline. The majority of measures had Child ICCs between .3 and .5. 7 measures had Child ICCs higher or equal to .5. Surprisingly, the top 6 metrics in terms of Child ICC corresponded to the “other child” category, known to have the worst accuracy according to previous analyses (Cristia et al., 2020). In an analysis fully reported in supplementary materials (SM K), we find some evidence that this may be due to the presence versus absence of siblings. The next metric with the highest Child ICC corresponded to an output measure, namely the total vocalization duration per hour extracted from ACLEW annotations (`voc_dur_chi_ph`, `aclew`), with a Child ICC of 0.5. Among adult metrics, the average vocalization duration for female vocalizations for ACLEW (`avg_voc_dur_fem`, `aclew`) and the ACLEW equivalent of CTC had the highest Child ICC (0.45 and 0.46, respectively).

### SM K: Are high Child ICCs for “other child” measures due to number or presence of siblings? (Exploration)

We reasoned the high Child ICC for metrics related to other children may be because children in our corpora vary in terms of the number of siblings they have, that siblings’ presence may be stable across recordings, and that a greater number of siblings would lead to more other child vocalizations. As a result,

any measure based on other child vocalizations would result in stable relative scores of children due to the number of siblings present. To test this hypothesis, we selected the metric with the highest Child ICC, namely ACLEW’s total vocalization duration by other children. We fit the full model again to predict this metric, but this time, in addition to controlling for age, we included sibling number as a fixed effect  $lmer(\text{metric} \sim \text{age} + \text{sibling\_number} + (1|\text{corpus/child}))$ , so that individual variation that was actually due to sibling number was captured by that fixed effect instead of the random effect for child. We had sibling number data for 874 recordings from 120 in 5 corpora (bergelson, lucid, quechua, warlaumont, winnipeg). The number of siblings varied from 0 to 7, with a mean of 0.9 and a median of 1. Results indicated the full model was singular, so we fitted a No Corpus model to be able to extract a Child ICC. As a sanity check, we verified that the number of siblings predicted the outcome, total vocalization duration by other children – and found that it did:  $\beta = 0.2$ ,  $t = 4.73$ ,  $p < .001$ . This effect is relatively small: It means that per additional sibling, there is a .2 standard deviation increase in this variable. Turning now to how much variance is allocated to the random factor of Child, there was no difference in Child ICC in our original analysis (0.64) versus this re-analysis including the number of siblings (0.64).

Perhaps it is not so much the sheer number of siblings that explains variance, but the sheer presence versus absence. After all, we can imagine that the effect of the number of siblings is not monotonic. We therefore repeated the analysis above but rather than adding the actual number of siblings, we had a binary variable that was “present” if the child had any siblings, and “absent” otherwise.

As in the sibling number analysis, the full model was singular, so we fitted a No Corpus model to be able to extract a Child ICC. We again verified that sibling presence predicted the outcome, total vocalization duration by other children – and found that it did:  $\beta = 0.97$ ,  $t = 9.91$ ,  $p < .001$ . This effect is, as expected, sizable: It means that there is nearly one whole standard deviation increase in this variable when there are any siblings. In addition to being a better predictor, in this model, the amount of variance allocated to individual children as measured by Child ICC was considerably higher in our original analysis (0.64) than in this re-analysis including sibling presence (0.52).

## SM L: Are “bad” output measures those coming from VCM? (Exploration)

Among ACLEW measures, a fair number of them come from VCM, a module that classifies child vocalizations in terms of vocal maturity types into cry, canonical, and non-canonical categories. In unpublished analyses, we have found that VCM labels are inaccurate when compared to human labels of the same vocalizations, relatively to other metrics. In this analysis, we checked whether VCM-derived measures had lower Child ICC than other ACLEW measures. As shown in the next Figure, this was not the case: Some output measures from the ACLEW pipeline have lower Child ICC than VCM ones.

## SM M: Code to reproduce Figure 5

Please see code in the RMarkdown version of the document.

## SM N: Code to reproduce text below Figure 5

Please see code in the RMarkdown version of the document.

Next, we explored how similar Child ICCs were across different talker types and pipelines. We fit a linear model with the formula  $lm(\text{icc\_child\_id} \sim \text{type} * \text{pipeline})$ , where type indicates whether the measure pertained to the key child, (female/male) adults, other children; and pipeline LENA or ACLEW. The model was overall significant ( $F(65) = 12.11$ ,  $p < .001$ ). We found an adjusted R-squared of 57%, suggesting much of the variance across Child ICCs was explained by these factors. A Type 3 ANOVA on this model revealed type was a significant predictor ( $F(4) = 26$ ,  $p < .001$ ), as was pipeline ( $F(1) = 2.8$ ,  $p = 0.097$ ); the interaction between type and pipeline was not significant. The main effect of type emerged because output metrics tended to have higher Child ICC ( $M = .4$ ,  $SD = .06$ ) than those associated to adults in general ( $M = .31$ ,  $SD = .08$ ), females ( $M = .34$ ,  $SD = .07$ ), and males ( $M = .35$ ,  $SD = .05$ ); whereas those associated with other children had even higher Child ICCs ( $M = .58$ ,  $SD = .06$ ). The main effect of pipeline arose



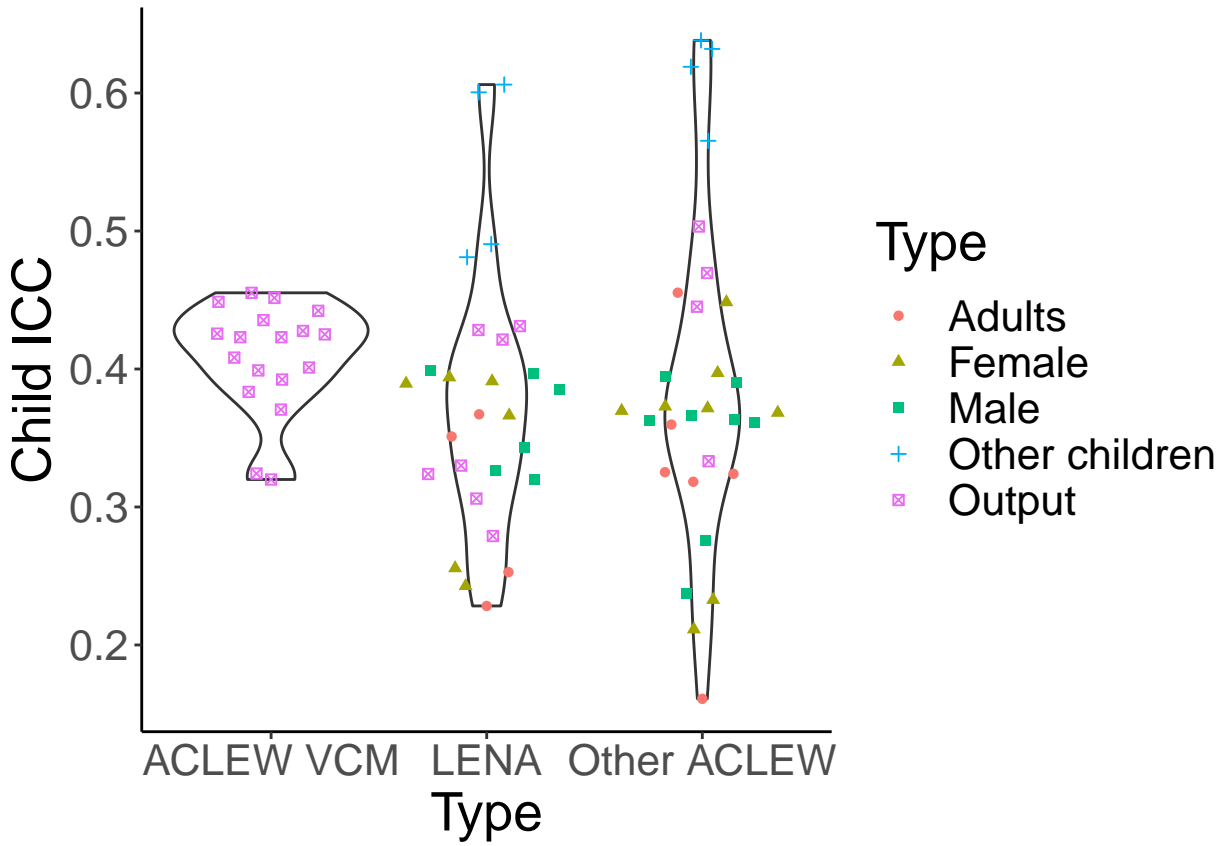


Figure 4: Figure SM L. Violin plot reflecting the distribution of Child ICC for ACLEW VCM versus other ACLEW or LENA metrics.

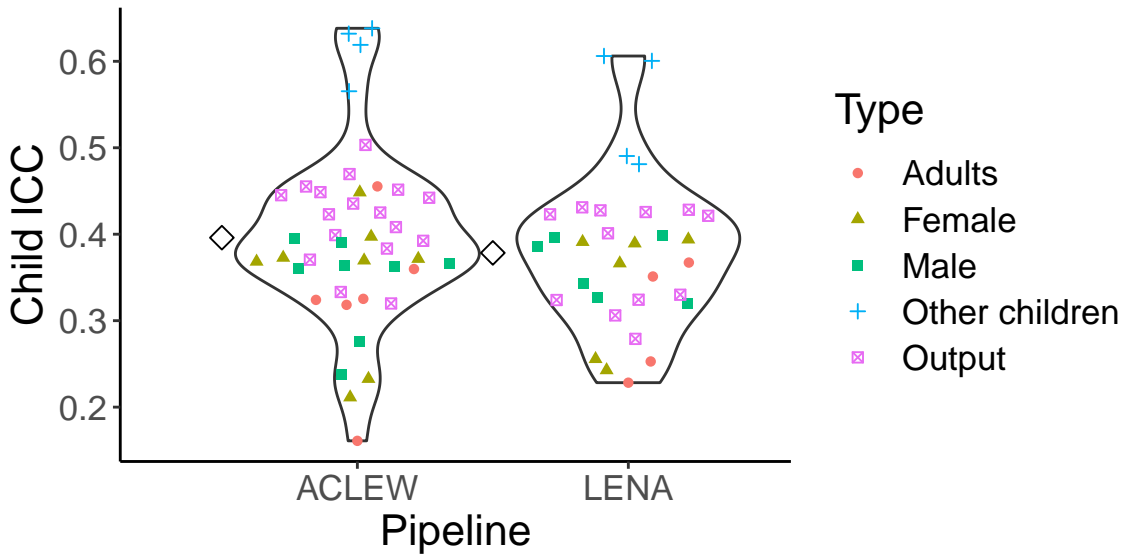


Figure 5: Figure 5 (reproduced). Violin plot reflecting the distribution of Child ICC.

Table 4: Table 4 (reproduced). Most commonly used metrics.

metric	LENA ICC	ACLEW ICC
wc_adu_ph	0.37	0.33
lena_CVC_ph	0.43	0.42
lena_CTC_ph	0.35	0.46
voc_fem_ph	0.39	0.37
voc_chi_ph	0.42	0.45

because of slightly higher Child ICCs for the ACLEW metrics ( $M = .4$ ,  $SD = .1$ ) than for LENA metrics ( $M = .38$ ,  $SD = .09$ ).

### SM O: Code to reproduce Table 4

Please see code in the RMarkdown version of the document.

### SM P: Code to reproduce text at the beginning of the “Reliability across age groups” section

Please see code in the RMarkdown version of the document.

Out of 450 fitted models (75 metrics times 6 age bins), 6 were singular when including a random intercept per child, and therefore they could not be included in these analyses at all. In addition, 146 were singular when including a random intercept per corpus. The remaining 298 could be analyzed with the full model.

### SM Q: Code to reproduce Figure 6

Please see code in the RMarkdown version of the document.

### SM R: Code to reproduce text below Figure 6

Please see code in the RMarkdown version of the document.

To interrogate these results statistically, and assess whether Child ICCs tended to be higher or lower in certain age bins, we fit a linear model with the formula  $lm(Child\_ICC \sim type * pipeline * age\_bin)$ . The model was overall significant ( $F(384) = 5.84$ ,  $p < .001$ ). We found an adjusted R-squared of 39%, suggesting this model explained more than a third of the variance in Child ICC. A Type 3 ANOVA on this model revealed that the interactions between type and pipeline, pipeline and age, and the three-way interaction (type, pipeline, age) were not significant. However, both the type by age bin interaction ( $F(20) = 6.3$ ,  $p < .001$ ) and the three main effects were significant (type:  $F(4) = 35.5$ ,  $p < .001$ ; age:  $F(5) = 8.5$ ,  $p < .001$ ; pipeline:  $F(1) = 8.5$ ,  $p = .01$ ).

See table below for results of the Type 3 ANOVA.

### SM S: Code to reproduce Figure 7

Please see code in the RMarkdown version of the document.

### SM T: Code to reproduce text in the “Reliability within corpus” section

Please see code in the RMarkdown version of the document.

Out of 600 fitted models (75 metrics times 8 corpora), 26 were singular when including a random intercept per child, and therefore they could not be included in these analyses at all. (Including a

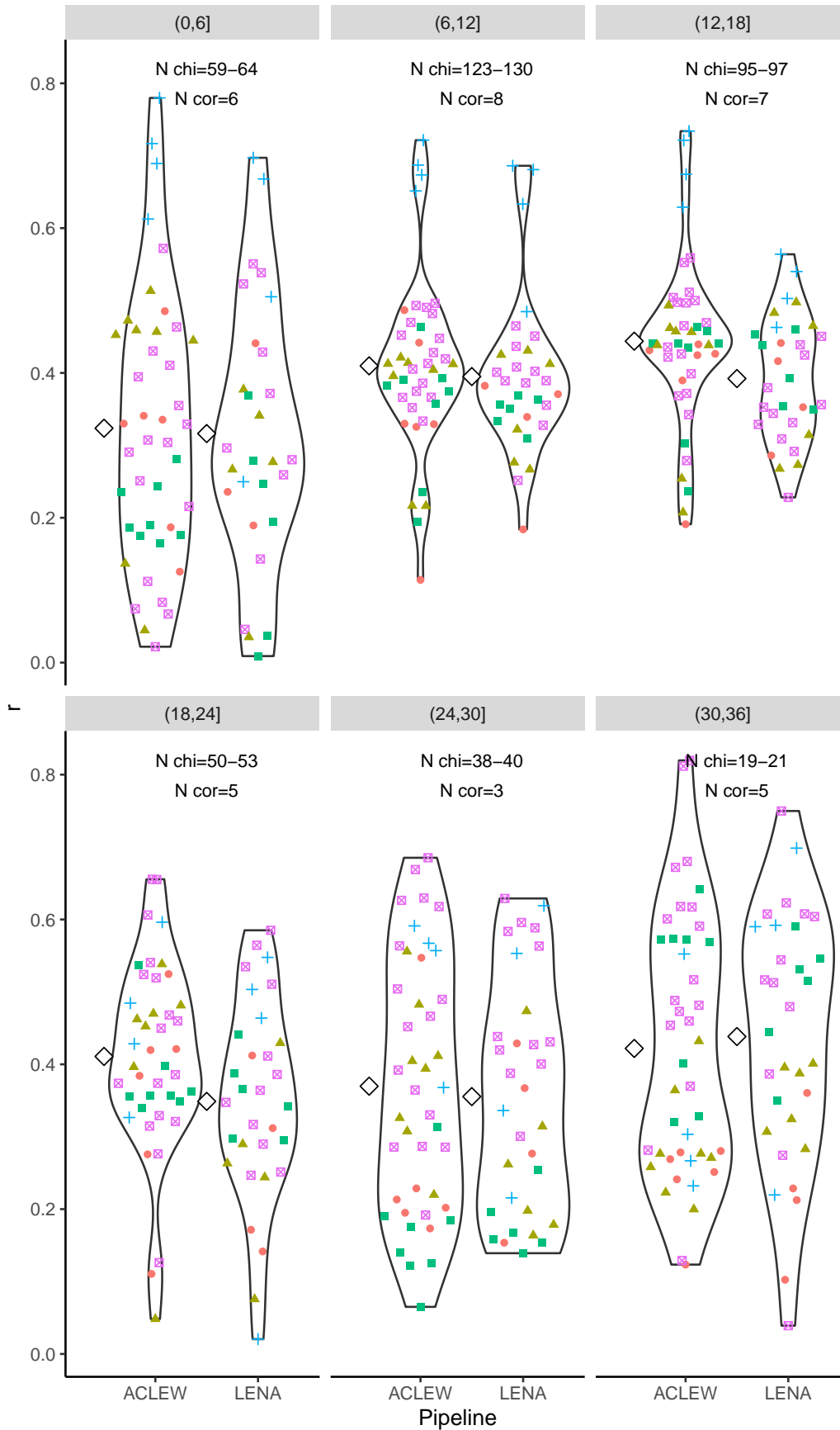


Figure 6: Figure 6 (reproduced). Child ICC by metric type and pipeline, when considering each age bin separately.

Table 5: Type 3 ANOVA on model attempting to explain variation in Child ICC as a function of talker types, pipelines, and age bins.

	Sum Sq	Df	F value	Pr(>F)
Type	2.05	4	35.51	0.00
data_set	0.08	1	5.79	0.02
age_bin	0.61	5	8.48	0.00
Type:data_set	0.05	4	0.78	0.54
Type:age_bin	1.81	20	6.28	0.00
data_set:age_bin	0.08	5	1.10	0.36
Type:data_set:age_bin	0.34	20	1.17	0.28
Residuals	5.54	384	NA	NA

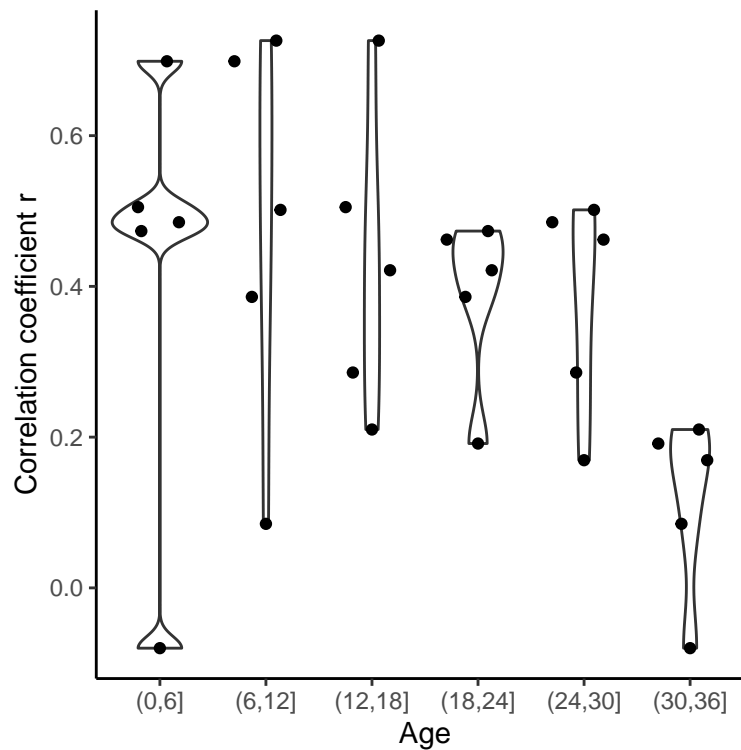


Figure 7: Figure 7 (reproduced). Correlations in Child ICC across age bins. Each point indicates the correlation in Child ICC for the age bin named in the x-axis with every other age bin.

random intercept per corpus is not relevant here, since only data from one corpus is included in each model fit.)

## SM U: Code to reproduce Figure 8

Please see code in the RMarkdown version of the document.

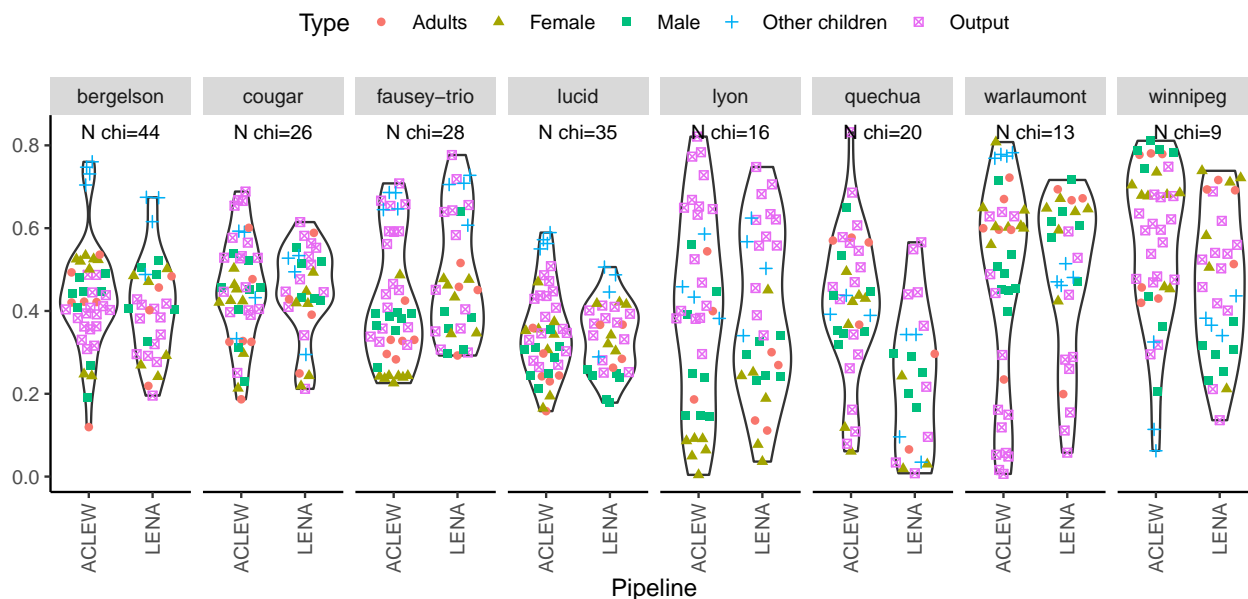


Figure 8: Figure 8 (reproduced). Child ICC by metric type and pipeline, when considering each corpus separately.

## SM V: Code to reproduce text below Figure 8

Please see code in the RMarkdown version of the document.

The fact that we cannot infer reliability from one corpus based on another one was confirmed statistically: We checked whether Child ICC differed by talker types and pipelines across corpora by fitting a linear model with the formula  $lm(Child_{ICC} \sim type * pipeline * corpus)$ , where type indicates whether the measure pertained to the key child, (female/male) adults, other children; pipeline LENA or ACLEW; and corpus the corpus ID. The model was overall significant ( $F(494) = 7.42, p < .001$ ). We found an adjusted R-squared of 47%, suggesting this model explained nearly half of the variance in Child ICC. A Type 3 ANOVA on this model revealed several significant effects and interactions, including a three-way interaction of type, pipeline, and corpus ( $F(28) = 2, p < .001$ ); a two-way interaction of type and corpus ( $F(7) = 4.9, p < .001$ ); and a main effect of corpus ( $F(7) = 16.5, p < .001$ ).

See Table below for results of the Type 3 ANOVA.

## SM W: Code to reproduce Figure 9

Please see code in the RMarkdown version of the document.

## SM X: Code to reproduce text in the Discussion section

Please see code in the RMarkdown version of the document.

Table 6: Type 3 ANOVA on model attempting to explain variation in Child ICC as a function of talker types and pipelines across corpora.

	Sum Sq	Df	F value	Pr(>F)
Type	0.77	4	11.58	0.00
data_set	0.12	1	7.46	0.01
corpus	1.92	7	16.54	0.00
Type:data_set	0.04	4	0.65	0.63
Type:corpus	5.37	28	11.55	0.00
data_set:corpus	0.57	7	4.93	0.00
Type:data_set:corpus	0.92	28	1.98	0.00
Residuals	8.20	494	NA	NA

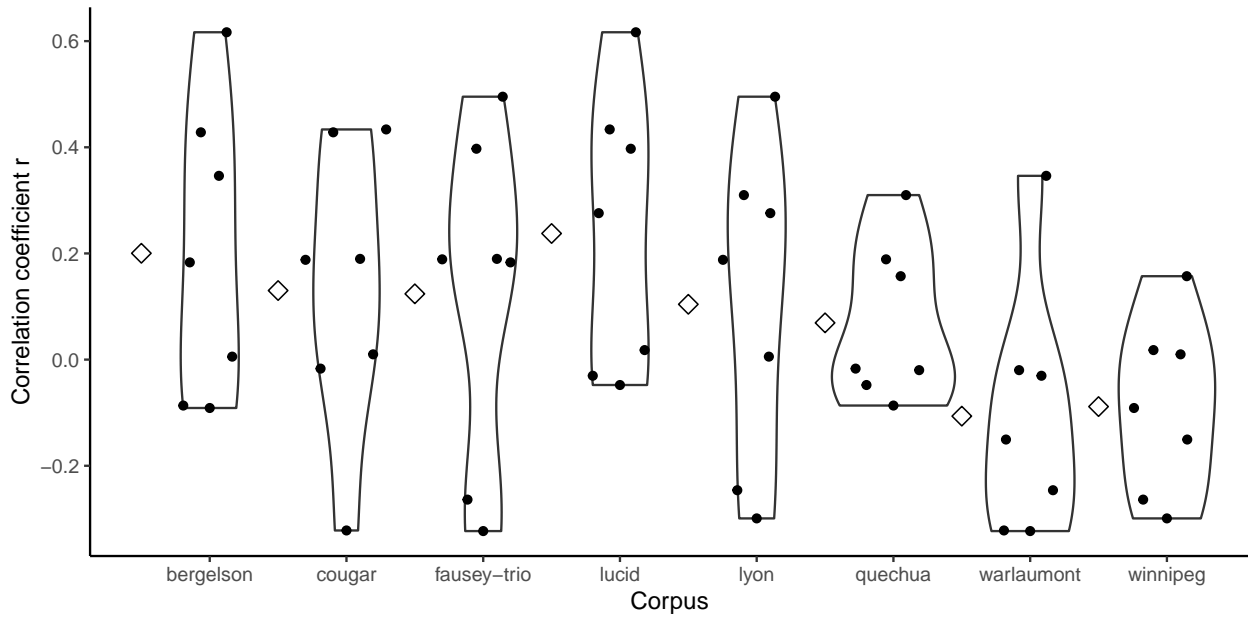


Figure 9: Figure 9 (reproduced). Correlations in Child ICC across corpora.

Our data draws mainly from urban (96% of recordings, 90% of the children, 88% of the corpora), English-speaking settings (92% of recordings, 81% of the children, 75% of the corpora), largely from North America (73% of recordings, 63% of the children, 62% of the corpora).

## SM Y: Variability as a function of hardware

Another potential negative contribution to reliability that is currently not discussed is variability in the experimental setup. In a corpus collected in the Solomon Islands, children were wearing two recorders simultaneously. These were USB devices, sourced from two different providers. In this dataset, the duration of the recordings could be very different within the same pair (a ~10% difference was not atypical), which means that what is actually recorded is somewhat random in itself. Even comparing identical audio ranges covered by both recordings of each pair, the corresponding ACLEW metrics differed slightly; they were strongly correlated ( $R^2$  was close to 0.95) but not perfectly correlated. This suggests that randomness in the recorders properties and their placement may also contribute to a decrease reliability. This reliability is importantly not at all due to changing underlying conditions, as both recorders picked up on the exact same day, so it is not due to variability in underlying behaviors. It is also not due to algorithmic variation because ACLEW algorithms are deterministic. Thus, this variability is only due to hardware differences and potentially also differences in e.g. USB placement.

## SM Z: References

- Alcock, K., Meints, K., & Rowland, C. (2020). *The UK Communicative Development Inventories*. London, UK: J&R Press.
- Cavallera, V., Lancaster, G., Gladstone, M., Black, M. M., McCray, G., Nizar, A., . . . & Janus, M. (2023). Protocol for validation of the Global Scales for Early Development (GSED) for children under 3 years of age in seven countries. *BMJ open*, 13(1), e062562. <http://dx.doi.org/10.1136/bmjopen-2022-062562>
- Denman, D., Speyer, R., Munro, N., Pearce, W. M., Chen, Y.-W., & Cordier, R. (2017). Psychometric Properties of Language Assessments for Children Aged 4–12 Years: A Systematic Review. *Frontiers in Psychology*, 8. <https://doi.org/10.3389/fpsyg.2017.01515>
- Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., & Pethick, S. J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, 59(5).
- Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and Consistency in Early Language Learning: The Wordbank Project*. Cambridge: MIT Press.
- Goldman, R., & Fristoe, M. (2015). *GFTA-3: Goldman Fristoe 3 Test of Articulation*. PsychCorp.
- Gurley, J. (2013). *Expressive Vocabulary Test, Second Edition*. In Goldstein, S. & Nagliery, J. A., *Encyclopedia of Child Behavior and Development*. Boston, MA: Springer US.
- Kobsar, D., Charlton, J. M., Tse, C. T. F., Esculier, J.-F., Graffos, A., Krowchuk, N. M., Thatcher, D., & Hunt, M. A. (2020). Validity and reliability of wearable inertial sensors in healthy adult walking: A systematic review and meta-analysis. *Journal of NeuroEngineering and Rehabilitation*, 17(1), 62. <https://doi.org/10.1186/s12984-020-00685-3>
- Kriener-Althen, K., Newton, E. K., Draney, K., & Mangione, P. L. (2020). Measuring readiness for kindergarten using the Desired Results Developmental Profile. *Early Education and Development*, 31(5), 739-763. <https://doi.org/10.1080/10409289.2020.1743160>
- Leaders Project (Nov 25, 2013). *Test Review: PSL-5 English*. Retrieved December 8, 2023, from <https://www.leadersproject.org/2013/11/25/test-review-pls-5-english/>.
- Liu, R., Zhang, R., Qu, Y., Jin, W., Dong, B., Liu, Y., & Mao, L. (2022). Reliability analysis of inertial sensors for testing static balance of 4-to-5-year-old preschoolers. *Gait & Posture*, 92, 176–180. <https://doi.org/10.1016/j.gaitpost.2021.11.029>

- NIH (2023). NIH Infant and Toddler (Baby) Toolbox. Retrieved March 16, 2023, from <https://neuroscience.blueprint.nih.gov/resources-tools/blueprint-resources-tools-library/nih-infant-and-toddler-baby-toolbox>
- Sahli, A. S., & Belgin, E. (2017). Adaptation, validity, and reliability of the Preschool Language Scale–Fifth Edition (PLS–5) in the Turkish context: The Turkish Preschool Language Scale–5 (TPLS–5). *International Journal of Pediatric Otorhinolaryngology*, 98, 143–149. <https://doi.org/10.1016/j.ijporl.2017.05.003>
- Schweinhart, L. J., Mcnair, S., Barnes, H., & Larner, A. M. (1993). Observing Young Children in Action to Assess their Development: The High/Scope Child Observation Record Study. *Educational and Psychological Measurement*, 53(2), 445–455. <https://doi.org/10.1177/0013164493053002014>
- Velikonja, T., Edbrooke-Childs, J., Calderon, A., Slead, M., Brown, A., & Deighton, J. (2017). The psychometric properties of the Ages & Stages Questionnaires for ages 2-2.5: A systematic review. *Child: Care, Health and Development*, 43(1), 1–17. <https://doi.org/10.1111/cch.12397>
- Williams, K. T. (1997). Expressive vocabulary test second edition (EVT™ 2). *J. Am. Acad. Child Adolesc. Psychiatry*, 42, 864-872.
- Zimmerman, I. L. , Steiner, V. G. , & Pond, R. E. (2011). Preschool Language Scales–5th Edition (PLS-5). Pearson. <https://doi.org/10.1037/t15141-000>